Data Wrangling with MongoDB
OpenStreetMap Project

Tianzhixi Yin

September 2015

1 Map Area

Lincoln, Nebraska, United States
https://www.openstreetmap.org/relation/169588

2 Problems Encountered in the Map

Street Names

After downloading the OSM file for Lincoln, Nebraska, I first ran it against AuditStreetName.py to find all the irregular street types. I noticed some abbreviations like ‘Ave’, ‘Dr’, and ‘St’. I updated them to the standard types, so ‘W Kearney Ave’ became ’W Kearney Avenue’ and ‘Pioneer Woods Dr’ became ‘Pioneer Woods Drive’ and so on. This was done before I imported the data to MongoDB.
The following corrections were done in MongoDB:
1. people ignored the street types when they typed the street names. So I changed ‘K’, ‘O’, and ‘P’ to ‘K Street’, ‘O Street’, and ‘P Street’. ‘South 10th’ was changed to ‘South 10th Street’, ‘North 12th’ was changed to ‘North 12th Street’, and ‘NW 1st’ was changed to ‘Northwest 1st Street’ (after making sure there are such street names in Lincoln).
2. people typed in a number after the street name. First I thought this was the house number mistyped here, but when I actually looked at all the records like this, house numbers were already there. I guess these are the room numbers that people don’t know where to put them. Therefor, for instance ‘NW 1st St 102’ was changed to ‘Northwest 1st Street’ and I would suggest ‘unit’ be created with the value ‘102’ (after making sure there is no original ‘unit’ input).

Country

I checked the ‘country’ part in addresses, all 60 inputs are correctly ‘US’.

State

I checked the ‘state’ part in addresses, all 1069 inputs are correctly ‘NE’.

City

I checked the ‘city’ part in addresses, the output is as follow:

```
{u’_id’: u’Lincoln’, u’count’: 914}
{u’_id’: u’Lincoln, NE’, u’count’: 3}
{u’_id’: u’Seward’, u’count’: 1}
{u’_id’: u’Denton’, u’count’: 1}
{u’_id’: u’Malcolm’, u’count’: 1}
```

The only problematic entries are the three ‘Lincoln, NE’s, I changed all to ‘Lincoln’ and updated the ‘state’ part to ‘NE’ (I checked and found all three were missing the ‘state’ input).

Postcode

I sorted the postcodes by count , part of the output was as follow:

```
{u’_id’: u’68508’, u’count’: 63}
{u’_id’: u’68504’, u’count’: 53}
{u’_id’: u’68462’, u’count’: 52}
......
{u’_id’: u’68526’, u’count’: 2}
{u’_id’: u’NE 68339’, u’count’: 1}
{u’_id’: u’68523’, u’count’: 1}
{u’_id’: u’68526-0736’, u’count’: 1}
{u’_id’: u’68509’, u’count’: 1}
......
```

The only problematic one is ‘NE 68339’, I changed it to ‘68339’ and updated the ‘state’ part to ‘NE’ (no state input originally).

House Number

All are numbers, no obvious problem.

3 Data Overview

This section contains basic statistics about the dataset and the MongoDB queries used to gather them.

File sizes
lincoln_nebraska.osm......... 67 MB

lincoln_nebraska.osm.json .... 74 MB

# Number of documents

```
> db.lincoln.find().count()
344308
```

# Number of nodes

```
>  db.lincoln.find({"type":"node"}).count()
309717
```

# Number of ways

```
> db.lincoln.find({"type":"way"}).count()
34589
```

# Number of unique users

```
> len(db.lincoln.distinct("created.user"))
198
```

# Top 1 contributing user

```
> result = db.lincoln.aggregate([{"$match":{"created.user":{"$exists":1}}},
                               {"$group":{"_id":"$created.user","count":{"$sum":1}}},
{"$sort":{"count":-1}},
{"$limit":1}])
for ele in result:
    pprint.pprint(ele)

{u'_id': u'Your Village Maps', u'count': 148613}
```

# Number of users appearing only once (having 1 post)

```
> result = db.lincoln.aggregate([{"$match":{"created.user":{"$exists":1}}},
                               {"$group":{"_id":"$created.user","count":{"$sum":1}}},
{"$group":{"_id":"$count","num_users":{"$sum":1}}},
{"$sort":{"_id":1}},
{"$limit":1}])
for ele in result:
    pprint.pprint(ele)

{u'_id': 1, u'num_users': 41}
```

4 Additional Ideas about the datasets

Other Address Information

About the numbers after street names, my best guess is that those are the room numbers, but I could be wrong, if we really want to correct those, we should confirm with information from other sources.

House Name

When I checked the house names, part of the output was:

```
{u'_id': u'The Sawmill Building', u'count': 3}
{u'_id': u'Armour Building', u'count': 2}
{u'_id': u'Lincoln Hide & Fur Bldg', u'count': 2}
{u'_id': u'Village of Bennet', u'count': 1}
{u'_id': u'Suite #204', u'count': 1}
{u'_id': u'The Courtyards', u'count': 1}
......
```

The only seemingly error is 'Suite 204', after checking the record I realized maybe the person just mistyped the room number into 'housename'. However, I don't think it is necessary to change this one. So I would recommend renaming the 'housename' section to 'buildingname' because people might want to put the room numbers in 'housename' and that could be confused with 'suite' or 'unit'.

Suite

Only one record found, no problem present.

Unit

No problem present, although some people seem to have confused 'unit' with 'suite', like the following example for 'unit':

```
{u'_id': u'D', u'count': 2}
{u'_id': u'Suite A', u'count': 1}
......
{u'_id': u'1,2', u'count': 1}
{u'_id': u'Suite J', u'count': 1}
......
```

But this is not a big problem.
Therefore, I would recommend replacing 'suite' and 'unit' with just 'roomnumber' because it seems like people are confused where to put the room number.

Contributor

The top 5 contributors have contributed 89.14% of the total data:

```
{u'_id': u'Your Village Maps', u'count': 148613}
{u'_id': u'behemoth14', u'count': 79833}
{u'_id': u'woodpeck_fixbot', u'count': 28289}
{u'_id': u'PHerison', u'count': 26407}
{u'_id': u'James GIS', u'count': 23786}
```

More than 40% of the users have only contributed 5 or less times. If we can find ways to encourage ordinary people to contribute, the map should be more detailed.

Additional data exploration using MongoDB queries

# Top 5 appearing amenities

```
> result = db.lincoln.aggregate([{"$match":{"amenity":{"$exists":1}}},
                                 {"$group":{"_id":"$amenity","count":{"$sum":1}}},
{"$sort":{"count":-1}},
{"$limit":5}])
for ele in result:
    pprint.pprint(ele)

{u’_id’: u’parking’, u’count’: 567}
{u’_id’: u’place_of_worship’, u’count’: 228}
{u’_id’: u’school’, u’count’: 138}
{u’_id’: u’fast_food’, u’count’: 137}
{u’_id’: u’restaurant’, u’count’: 104}
```

# Top 5 religions

```
> result = db.lincoln.aggregate([{"$match":{"amenity":{"$exists":1}, "amenity":"place_of_wor
                                 {"$group":{"_id":"$religion","count":{"$sum":1}}},
{"$sort":{"count":-1}},
{"$limit":5}])
for ele in result:
    pprint.pprint(ele)

{u’_id’: u’christian’, u’count’: 218}
{u’_id’: None, u’count’: 4}
{u’_id’: u’buddhist’, u’count’: 3}
{u’_id’: u’jewish’, u’count’: 1}
{u’_id’: u’muslim’, u’count’: 1}
```

# Top 5 denominations

```
> result = db.lincoln.aggregate([{"$match":{"amenity":{"$exists":1}, "amenity":"place_of_wor
                                 {"$group":{"_id":"$denomination","count":{"$sum":1}}},
{"$sort":{"count":-1}},
{"$limit":5}])
for ele in result:
    pprint.pprint(ele)

{u’_id’: None, u’count’: 119}
{u’_id’: u’lutheran’, u’count’: 30}
{u’_id’: u’methodist’, u’count’: 23}
{u’_id’: u’baptist’, u’count’: 18}
{u’_id’: u’catholic’, u’count’: 13}
```

Maybe many churches are non-denominational.

# Top 5 cuisines in restaurants

```
> result = db.lincoln.aggregate([{"$match":{"amenity":{"$exists":1}, "amenity":"restaurant"}
                               {"$group":{"_id":"$cuisine","count":{"$sum":1}}},
{"$sort":{"count":-1}},
{"$limit":5}])
for ele in result:
    pprint.pprint(ele)

{u'_id': u'american', u'count': 27}
{u'_id': None, u'count': 15}
{u'_id': u'pizza', u'count': 8}
{u'_id': u'chinese', u'count': 8}
{u'_id': u'italian', u'count': 5}
```

# Top 5 cuisines in fast foods

```
> result = db.lincoln.aggregate([{"$match":{"amenity":{"$exists":1}, "amenity":"fast_food"}}
                               {"$group":{"_id":"$cuisine","count":{"$sum":1}}},
{"$sort":{"count":-1}},
{"$limit":5}])
for ele in result:
    pprint.pprint(ele)

{u'_id': u'sandwich', u'count': 30}
{u'_id': u'burger', u'count': 24}
{u'_id': u'ice_cream', u'count': 16}
{u'_id': u'mexican', u'count': 15}
{u'_id': None, u'count': 14}
```

No surprise here.

5 Conclusion

I would say the data in Lincoln, NE are pretty well documented, not a lot of errors were found. The people who contributed to this area did a great job. However, if we want this map to be perfect, there are still some work to be done, especially some cleanings need extra information outside of OpenStreetMap to be made. Also, the data format could be improved a little by replacing the sections that are confusing for people. Overall, this is a fun project and I truly learned a lot.